Q-Learning: Off-Policy TD Control

- SARSA, as Policy Iteration in DP, is based on Bellman Expectation Equation

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right)$$

$$Q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \left(r + \gamma \sum_{a'} \pi(a'|s') Q_\pi(s',a')\right)$$

- Q-Learning, as Value Iteration in DP, is based on Bellman Optimality Equation

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right)$$

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a) \left(r + \gamma \max_{a'} Q^*(s',a')\right)$$